ISSN:0975-9646

A Survey on Privacy Preserving Data Mining Techniques

* Mayil.S and Vanitha.M

Research and Dept of Computer Science, J.J College , Pudukkottai,TamilNadu(St),India

***Abstract*— Enormous amount of detailed private data is recurrently collected and analysed by applications using data mining, sharing of these data is useful to the application users. While sharing the private data, privacy preserving is becoming an increasingly significant issue. Sequential pattern mining is the process of finding relevant pattern in the data set. Sequential pattern helps in envisaging the next event. Predicting the sequence datasets leads to violate the privacy and disclose sensitive patterns related to medical records, business secrets etc.**

This paper explores about different various techniques for privacy preserving data mining such as anonymity, randomization, secure multiparty computation, sequential pattern hiding.

***Keywords*— Privacy preserving; Sequential pattern mining; Anonymity; Randomization; Secure multiparty computation; Sequential pattern hiding**

I. INTRODUCTION

Data mining is a novel technique for intelligently extracting information or knowledge from a large amount of data [1]collected by individuals, governments, hospitals which has great opportunities to take out sensitive knowledge patterns [2].This has lead to better concern about the privacy of the personal information [3]. The complete data about an individual often includes some sensitive information. Distributing such data instantly violates individuals' privacy.

The concept of privacy preserving data mining involves in preserving personal information from data mining algorithms. Privacy preserving data mining technique [5] is a new research area in data mining and statistical databases where mining algorithms are analyzed for the side effect they acquire in data privacy. The objective of privacy preserving data mining is to build algorithms for transforming the original information in some way, so that the private data and private knowledge remain confidential even after the mining process [4].

Sequential pattern mining is other important computation. [7]Sequential pattern mining discovers sequence of patterns from the large database. The computation of mining patterns in sequence is specifically carried out over customer purchase behavior analysis in retailing business and medical record analysis [7]. The retailer can analyze the purchase behavior of customers to predict their needs and satisfy their demands. Under privacy limitations, the privacy preserving data mining problem was intensely researched. To solve this problem number of efficient techniques has been proposed for privacy preserving data mining. It can be done without compromising the security of user's data. But most of these methods might result with some drawbacks as information loss and side-effects to some extent. This paper presents a brief survey of different privacy preserving data mining techniques and analyses the specific methods for privacy preserving data mining.

II. PRIVACY PRESERVING TECHNIQUES

The main objective of privacy preserving data mining is to develop data mining methods without increasing the risk of mishandling [6] of the data used to generate those methods. Most of the techniques use some form of alteration on the original data in order to attain the privacy preservation. The altered dataset is obtainable for mining and must meet privacy requirements without losing the [6] benefit of mining.

A.Randomization

Randomization technique is an inexpensive and efficient approach for privacy preserving data mining (PPDM). In order to assure the performance [12] of data mining and to preserve individual privacy, this randomization schemes need to be implemented. The randomization approach protects the customers' data by letting them arbitrarily alter their records before sharing them, taking away some true information and

and introducing some noise. Some methods in randomization are numerical randomization and item set randomization Noise can be introduced either by adding or multiplying random values to numerical records (Agrawal&Srikant, 2000) or by deleting real items and adding "fake" values to the set of attributes.

B. *Anonymization*

To protect individuals' identity when releasing sensitive information, data holders often encrypt or remove explicit identifiers, such as names and unique security numbers. However, unencrypted data provides no guarantee for anonymity. In order to preserve privacy, k-anonymity model has been proposed by Sweeney [8] which achieves k-anonymity using generalization and suppression [6], In K-anonymity, it is difficult for an imposter to determine the identity of the individuals in collection of data set containing personal information. Each release of data contains every combination of values of quasi-identifiers and that is indistinctly matched to at least k-1 respondents [16]. Generalization involves replacing a value with a less specific (generalized) but semantically reliable value. For example, the age of the person could be generalized to a range such as youth, middle age and adult without specifying appropriately, so as to reduce the risk of identification. [6] Suppression involves reduce the exactness of applications and it does not liberate any information .By using this method it reduces the risk of detecting exact information.

C. *Secure multi-party computation*

An alternative approach based on the multiparty computation is that every part of private data is validly known to one or more parties. Revealing private data to parties such as by whom the data is owned or the individual to whom the data refers to is not a condition of violating privacy. The problem arises when the private information is revealed to some other third parties. To deal with this problem, we use a specialized form of privacy preserving distributed data mining. Parties that each knows some of the private data participate in a protocol that generates the data mining results, [11] that guarantees no data items is revealed to other parties. Thus the process of data mining doesn't cause, or even increase the opportunity for breach of privacy.

D. *Sequential pattern hiding*

Sequential pattern hiding method [17] is necessary to conceal sensitive patterns that can otherwise be extracted from published data, without seriously affecting the data and the non sensitive interesting patterns. [13]Sequential pattern hiding is a challenging problem, because sequences have more composite semantics than item sets, and calls for efficient solutions that offer high utility.

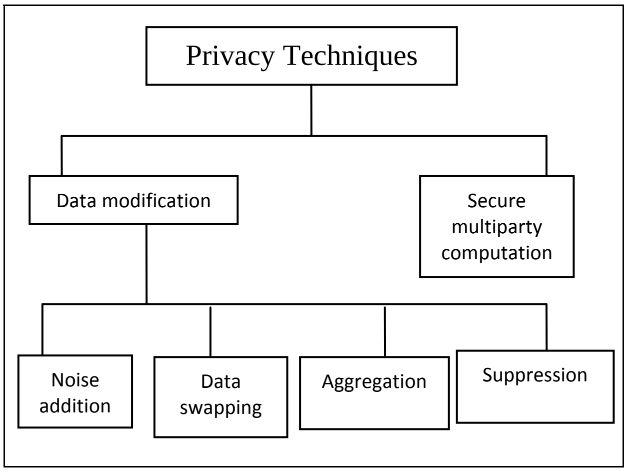

Fig.1 Classification of Privacy preserving techniques

III. RELATED WORK

This section briefly describes about the various techniques implemented in privacy preserving data mining.In the first method, Alexandre Evfimievski [10], chosen the randomization algorithm, so that collective properties of the data can be recovered with sufficient accuracy, while each entries are considerably indistinct. Here privacy measures were used to determine the distortion needed to protect privacy. This paper presents some randomization methods for kinds of data, and discusses the issue of measuring privacy.

In order to guarantee the performance of data mining and for the security of individual privacy, optimal randomization method needs to be implemented. Yu Zhu, Lei Liu [12] demonstrates the construction of optimal randomization schemes for privacy preserving density estimation. The effect of randomization on data mini v ng is computed by performance degradation and mutual information loss, while interval-based metrics are computed by privacy and privacy loss.

In this method, V. Ciriani, S. De Capitani di, Vimercati, S. Foresti, and P. Samarati [18], used k-anonymity to disclose the identity of individuals in the collection of dataset which is indistinctly matched to at least k-1 respondents. It measures the amount of anonymity retained during data mining. K- Anonymization method reduces the effectiveness of data mining algorithm on anonymized data and renders privacy preservation. While releasing truthful information, the original *k*-anonymity proposal and its enforcement via generalization and suppression to protect respondents' identities were illustrated and also discussed in different ways for applying generalization and suppression.

In this paper, Yehuda Lindell, Benny Pinkasy [19], presents introduction to secure multiparty computation and its applicability to privacy-preserving data mining. The common errors that are established in the literature when privacy-preserving data mining is implemented with secure multiparty computation techniques and the issues involved in the efficiency are discussed and also demonstrates the difficulties in constructing highly efficient protocols.

Ueli Maurer [20], proposed a simple approach to multi-party computation with straight-forward security proofs. This work achieves security only for a passive adversary setting, without the possibility to enhance it to active adversary settings. Due to their simplicity, the described protocols are well-suited for educational purposes, which is a main goal of this paper. Another advantage of the protocol used in this paper is it works over any field particularly over binary field.

In this method, Aris Gkoulalas-Divanis and Grigorios Loukides [13], proposed about sequential pattern hiding. Publishing sequence datasets offers remarkable opportunities for discovering interesting knowledge patterns. This paper considers how to clean the data to prevent the revealing of sensitive patterns during sequential pattern mining, while ensuring that the nonsensitive patterns can still be discovered. The first algorithm used in this paper attempts to sanitize data with minimal alteration, whereas the second focuses on minimizing the other side-effects.

In this paper Amruta Mhatre and Durga Toshniwal [14], presented a novel technique to hide sensitive co-occurring sequential patterns. This method works on progressive databases however, most of the standard techniques for privacy preservation work only on static database. Progressive databases are a generalized model of static, active and incremental [14]databases. The method is also extended to suit these different types of databases. The approach presented in this paper avoids the occurrence of most sensitive patterns frequently by suppressing [14] the patterns and keeping it from being frequent. It is further examined in order to develop methods to opt the pattern to be blocked.

Shikha Sharma & Pooja Jain [15], work is based on the reduction of support and confidence of sensitive rules. In this work algorithm is used in some modified form to hide the sensitive association rule without any side effect. To hide the sensitive element, algorithm repeatedly increases the hiding counter of the rule until confidence goes below a minimum specified threshold [15] rather than checking all transactions again and again and ordering them in increasing or decreasing order. If the confidence goes below minimum specified confidence threshold [15], rule is hidden i.e. it will not be discovered through any data mining algorithm.

IV. Conclusion

This paper presents a brief survey on various standard techniques for privacy preserving data mining was presented namely: randomization, anonymization, secure multiparty computation. Because of the increasing capability to trace and gather large amount of sensitive information, privacy preserving in data mining applications has become an important concern. Also, I have presented an overview about the sequential pattern mining which is the most effective area in data mining and for preserving privacy in extracting sequence of knowledge; sequential pattern hiding method is also discussed.

References

[1] Han Jiawei, M. Kamber, "*Data Mining: Concepts and Techniques*", Beijing: China Machine Press, pp.140, 2006.

[2] Benjamin C. M. Fung, Ke Wang, Rui Chen, Philip S. Yu, "*Privacy-Preserving Data Publishing: A Survey of Recent Developments*", ACM Computing Surveys, Vol. 42, No. 4, Article 14, Publication date: June 2010.

[3] Charu C. Aggarwal, Philip S. Yu, "*A General Survey of Privacy-Preserving Data Mining Models and Algorithm's*".

[4] Alexandre Evfimievski, Tyrone Grandison, "*Privacy Preserving Data Mining*".

[5] Evfimievski, A., R. Srikant, R. Agrawal and J. Gehrke. "*Privacy preserving mining of association rules*". Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-25, ACM Press,Edmonton, AB., Canada, pp. 1-12,2002.

[6] Pingshui WANG," *Survey on Privacy Preserving Data Mining*", International Journal of Digital Content Technology and its Applications, Volume 4, Number 9, December 2010.

[7] Chetna Chand, Amit Thakkar, Amit Ganatra, "*Sequential Pattern Mining: Survey and Current Research Challenges*",International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012.

[8] L. Sweeney, "*K-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*", 10 (5), 2002.

[9] Yogendra Kumar Jain, Vinod Kumar Yadav& Geetika S. Panday, *An Efficient Association Rule Hiding Algorithm for Privacy Preserving Data Mining*, International Journal on Computer Science and Engineering (IJCSE), Vol. 3 No. 7 July 2011.

[10] Alexandre Evfimievski, "*Randomization in Privacy Preserving Data Mining*", SIGKDD Explorations. Volume -4, Issue - 2, 2002.

[11] Jaideep Vaidya & Chris Clifton, "*Privacy-Preserving Data Mining: Why, How, and When*", the IEEE computer society, 2004.

[12] Yu Zhu& Lei Liu, "*Optimal Randomization for Privacy Preserving Data Mining*", ACM, August 2004.

[13] Aris Gkoulalas-Divanis, & Grigorios Loukides, "*Revisiting Sequential Pattern Hiding to Enhance Utility*", ACM, August 2011.